Neighborhood Effects in Integrated Social Policies Supplemental Appendix

Matteo Bobba* Jérémie Gignoux†

S1 Randomness of the Evaluation Sample

We hereby provide the english translation for a few extracts of an official document from the original evaluation of the program (Progresa, 1997), which clearly suggest that the evaluation sample was indeed selected randomly among the set of the program eligible localities in the seven Mexican States in which the program was initially implemented.

> [...]The evaluation sample (cf. BASAL y CONTROL) is constituted of rural (i.e. 50-4,999 inhabitants) and marginalized (i.e. high or very high values of an underlying proxy-mean poverty score) localities with access to primary and secondary schools that are located in *Progresa* catchment areas in the States of *Guerrero*, *Hidalgo*, *Michoacán*, *Puebla*, *Querétaro*, *San Luis Potosí* and *Veracruz*.
>
> [...]The sample was randomly drawn from the population of those program eligible localities scheduled to be incorporated in 1997, after stratification by geographic region (which roughly coincide with States) and population size.

In the regression analysis discussed in the paper, individual school participation decisions after the program takes place (1998-1999) are affected by the local frequency of program incorporated villages over the same period in the areas surroundings their villages of residence. Hence, beyond the presumed randomness of the evaluation sample vis-a-vis the villages that were incorporated during the baseline of the program evaluation (1997), we are also interested in assessing the extent to which the evaluation villages are similar to those that were incorporated in the program during the subsequent phases of the roll-out of the intervention in rural areas. Basic socio-demographic variables extracted from the 2000 Mexican population census indicate that localities that are lately incorporated into the program are on average larger and less marginalized (see Panel A of Table S1.1). However, the differences in means between incorporation phases are largely attenuated once we restrict the sample to those localities that are situated in 5km neighborhoods (as defined in the paper) of the evaluation localities (see panel B of Table S1.1).

*Toulouse School of Economics, University of Toulouse Capitole, 21 Allée de Brienne 31000 Toulouse France. E-mail: matteo.bobba@tse-fr.eu.

†French National Institute for Agricultural Research and Paris School of Economics, 48 boulevard Jourdan 75014 Paris France. E-mail: jeremie.gignoux@psemail.eu.

Table S1.1: Socio-Demographics Characteristics and the Phase of Program Incorporation

Year	1997		1998		1999	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: All Localities in the Seven States in which the Program Evaluation Took Place						
Size size	N=2,249		N=11,987		N=6,124	
Poverty mean-score	0.63	0.65	0.61	0.69	0.09	0.88
Nb of Households	70.17	94.71	73.74	99.91	91.35	171.17
Population (age≤5)	59.58	83.63	63.99	88.98	71.93	141.81
Population (6≤age<15)	91.47	124.62	96.64	129.09	108.20	204.99
Population (age≥15)	202.9	280.1	207.5	279.6	258.1	489.5
Presence of secondary school	0.20	0.40	0.19	0.40	0.22	0.42
% Literate (age≥15)	0.69	0.13	0.70	0.14	0.75	0.14
% Children in School (6≤age<15)	0.90	0.09	0.88	0.10	0.87	0.12
Altitude (meters above see level)	940.2	683.2	1092.7	855.6	1238.5	886.6
% Population in Workforce	0.43	0.11	0.40	0.13	0.39	0.13
Panel B: Only Localities in a 5-km Neighborhood of the Evaluation Sample						
Size size	N=717		N=3,048		N=829	
Poverty mean-score	0.66	0.64	0.67	0.72	0.16	0.99
Nb of Households	68.32	87.73	63.67	74.21	76.63	200.37
Population (age≤5)	57.78	78.32	55.83	68.24	60.02	149.05
Population (6≤age<15)	90.12	117.9	84.27	97.85	91.07	214.4
Population (age≥15)	198.3	261.4	180.9	211.1	216.0	568.6
Presence of secondary school	0.19	0.40	0.17	0.37	0.20	0.40
% Literate (age≥15)	0.68	0.13	0.68	0.15	0.72	0.17
% Children in School (6≤age<15)	0.90	0.10	0.88	0.10	0.88	0.13
Altitude (meters above see-level)	900.1	682.8	1215.9	842.9	1448.8	833.4
% Population in Workforce	0.44	0.12	0.42	0.14	0.42	0.14

NOTE: Both Samples of Panel A and Panel B include the 320 evaluation localities that were randomly assigned to the treatment group.

Table S1.2 provides a direct comparison between evaluation localities and the non-evaluation program beneficiary localities in the seven Mexican States from which the evaluation sample was drawn and over the period 1997-1999 (see columns 1 and 2). Column 3 reports the t-statistics of the test of no differences in locality characteristics after controlling for population strata and State fixed effects (as well as a joint F-test) and finds the presence of some unbalances in a few socio-demographic characteristics. In Column 4 we further restrict the comparison within the neighboring localities of each evaluation cluster and do not find evidence of any significant difference between the two samples.

To wrap up, in spite of the random assignment of evaluation villages in 1997, there may be some unbalancedness between the evaluation sample and the non-evaluation sample

of *Progresa* localities after 1997. However, those seem to be minor within the relatively homogenous group of neighboring villages that are incorporated in the program over the period 1998-1999 that we consider in our empirical analysis.

Table S1.2: Comparison of Means between Evaluation and Non-Evaluation *Progresa* Localities

Sample size	Evaluation N=506 (1)	Non-Evaluation N=20,045 (2)	T-test of No Difference N=20,551 (3)	T-test of No Difference N=4,785 (4)
Poverty mean-score	0.47	0.46	0.002	-0.039
	(0.73)	(0.79)	[0.971]	[0.112]
Number of Households	52.18	79.09	1.875	1.758
	(35.20)	(126.4)	[0.382]	[0.164]
Population (age≤5)	43.61	66.27	2.459	2.332
	(34.89)	(107.9)	[0.079]	[0.063]
Population (6≤age<15)	66.81	100.10	3.333	2.086
	(51.70)	(156.6)	[0.057]	[0.217]
Population (age≥15)	147.78	223.49	4.963	5.090
	(101.8)	(359.3)	[0.347]	[0.142]
Presence of Secondary School	0.17	0.20	0.024	0.011
	(0.38)	(0.40)	[0.453]	[0.516]
% Literate (age≥15)	0.71	0.71	-0.013	0.005
	(0.14)	(0.14)	[0.221]	[0.263]
% Children in School (6≤age<15)	0.88	0.88	-0.003	-0.002
	(0.10)	(0.11)	[0.666]	[0.612]
Altitude (meters above see level)	1273.63	1116.90	132.739	11.998
	(839.7)	(852.9)	[0.060]	[0.153]
% Population in Workforce	0.40	0.40	-0.002	0.001
	(0.13)	(0.13)	[0.792]	[0.796]
F Test of Joint Orthogonality			4.375	1.045
[p-values]			[0.048]	[0.404]

NOTE: Columns 1-2 report means and standard deviations (in parenthesis). Columns 3-4 display the OLS coefficients with State fixed effect (Column 3) and Neighborhood fixed effects (Column 4) of the evaluation dummy along with the p-values (in brackets) for the null hypothesis of no difference between evaluation and non-evaluation program localities. Population strata are included in both specifications. Standard errors are clustered at the state level in Column 3 and at the neighborhood level in Column 4.

Turning now to the sample used in the empirical analysis; a direct implication of the presumed randomness of the evaluation sample is that we should expect no differences in the survey characteristics measured in the pre-program year (1997) between evaluation neigh-

borhoods with a different number of non-centroid evaluation villages after netting out the independent effect of the local frequency of neighboring non-evaluation localities. A simple comparison of means between evaluation neighborhoods with and without non-centroid evaluation villages is shown in columns 1 and 2 of Table S1.3. Column 3 reports the effects of the number of evaluation villages after controlling for the number of non-evaluation villages in the neighborhood. Overall, the local frequency of evaluation localities in the neighborhoods of the 506 localities that form part to the *Progresa* evaluation sample does not seem to be correlated with pre-program observable characteristics at the individual, locality and neighborhood-level. This confirms that the set of evaluation localities is a random subsample of beneficiary localities incorporated after late 1997.

Table S1.3: Baseline Characteristics within the Evaluation Localities

Neighborhood	No Evaluation Localities (1)	Some Evaluation Localities (2)	OLS Coefficient of Number of Eval ($N^E_{j,5}$) (3)
School Enrollment	0.64	0.64	0.004
	(0.48)	(0.48)	[0.647]
Individual and HH Characteristics			
Age	14.52	14.53	-0.016
	(2.02)	(2.06)	[0.513]
Female (dummy)	0.50	0.51	0.002
	(0.50)	(0.50)	[0.729]
Mother Education (years)	2.36	2.36	0.077
	(2.33)	(2.35)	[0.396]
Father Education (years)	2.20	2.25	0.085
	(2.21)	(2.30)	[0.139]
Centroid Village Characteristics			
Share of Program Eligible HHs	0.58	0.60	-0.008
	(0.20)	(0.19)	[0.409]
Presence of Secondary School	0.26	0.24	-0.006
	(0.44)	(0.43)	[0.789]
Distance to Nearest City (Km)	106.4	102.9	4.191
	(40.45)	(45.59)	[0.255]
Neighborhood (radius=5km) Characteristics			
Number of Secondary Schools	2.88	3.19	0.037
	(2.19)	(1.91)	[0.800]
Poverty mean-score	0.41	0.57	0.042
	(0.56)	(0.54)	[0.151]
Number of Localities (any)	21.03	25.15	-0.181
	(14.00)	(11.21)	[0.798]
Population density	5.99	8.41	0.572
	(6.54)	(11.44)	[0.575]
F Test of Joint Orthogonality			1.099
[p-values]			[0.357]

NOTE: Columns 1-2 report means and standard deviations (in parenthesis). Column 3 displays the OLS coefficients of the number of neighboring (within 5km) evaluation localities after controlling for the number of neighboring (within 5km) non-evaluation *Progresa* localities along with the associated p-values (in brackets). State fixed effects are included in all specifications. Standard errors are clustered at the level of groupings of partially overlapping neighborhoods.

S2 Consistency of the OLS coefficients When One or More Regressor is Endogenous

A general writing of the linear regression model we consider is

$$y = x_1\beta_1 + x_2\beta_2 + u$$

where x_2 is a matrix of endogenous explanatory variables, i.e. $E(x_2'u) \neq 0$ and x_1 is a matrix of explanatory variables that are exogenous, i.e. i.e. $E(x_1'u) = 0$. β_1 and β_2 are (column) vectors of coefficients and u is a vector of the error terms.

In our specific case, x_2 is the number of potential beneficiary villages N_P, while x_1 is the number of actual beneficiary villages N_B. While N_B is correlated with treatment density and hence potentially associated with unobserved determinants of outcomes captured by the residual u in an unconditional model, it is not anymore when we control for N_P. This is because the residual variation in N^B is then solely determined by the random number of treatment group villages in the neighborhood. The model we estimate is formally similar to the general model above, as the same estimates can be obtained using the variables N^C and N^P, and N^C is exogenous while N^P is endogenous; we indeed have:

$$y = \beta_1 N^B + \beta_2 N^P + u = \beta_1(N^P - N^C) + \beta_2 N^P + u = -\beta_1 N^C + (\beta_1 + \beta_2)N^P + u$$

Using the partitioned matrix notation, the vector of OLS coefficients is given by

$$\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}x_1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2\end{pmatrix}^{-1}\begin{pmatrix}x_1' \\ x_2'\end{pmatrix} u$$

Since x_2 is random conditional on x_1, it follows that the two set of explanatory variables are uncorrelated with each others: $E(x_1'x_2) = 0$. Hence, the variance $V(x_1, x_2)$ of the full set of covariates in the right-hand side term converges asymptotically to

$$\begin{pmatrix}x_1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2\end{pmatrix} \to \begin{pmatrix}E(x_1'x_1) & 0 \\ 0 & E(x_2'x_2)\end{pmatrix}$$

Using the independence of x_2 from u, the product of the last two terms converges to

$$\begin{pmatrix}x_1' \\ x_2'\end{pmatrix} u \to \begin{pmatrix}E(x_1'u) \\ 0\end{pmatrix}$$

So the vector of coefficients converges asymptotically to

$$\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} \to \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} + \begin{pmatrix}E(x_1'x_1)^{-1}E(x_1'u) \\ 0\end{pmatrix}$$

To summarize, the OLS coefficients $\hat{\beta}_2$ are consistent estimates of the marginal effects of the exogenous variables, β_2. On the other hand, the OLS coefficients on the endogenous variables $\hat{\beta}_1$ are biased, due for instance to omitted variables, but this bias does not contaminate the

coefficients estimated for the exogenous variables to the extent that those are uncorrelated with the endogenous variables. In our setting, this is likely the case because the partial variation in N^B (i.e. after netting out the effect of N^P) solely captures the randomized allocation into treatment of evaluation localities.

Additional References

Progresa (1997). "Nota Tecnica: Diseño Muestra Basal y Control", *Technical Note*.